

TRANSFORMATIONS TO ADDITIVITY IN MEASUREMENT ERROR MODELS

September 1, 1996

R. Stephen Eckert
Lilly Research Laboratories
Eli Lilly and Company
Indianapolis IN 46285

Raymond J. Carroll and Naisyin Wang
Department of Statistics
Texas A&M University
College Station TX 77843-3143

SUMMARY

In many problems one wants to model the relationship between a response Y and a covariate X . Sometimes it is difficult, expensive, or even impossible to observe X directly, but one can instead observe a substitute variable W which is easier to obtain. By far the most common model for the relationship between the actual covariate of interest X and the substitute W is $W = X + U$, where the variable U represents measurement error. This assumption of *additive* measurement error may be unreasonable for certain data sets. We propose a new model, namely $h(W) = h(X) + U$, where $h(\cdot)$ is a monotone transformation function selected from some family \mathcal{H} of monotone functions. The idea of the new model is that, *in the correct scale*, measurement error is additive. We propose two possible transformation families \mathcal{H} . One is based on selecting a transformation which makes the within sample mean and standard deviation of replicated W 's uncorrelated. The second is based on selecting the transformation so that the errors (U 's) fit a prespecified distribution. Transformation families used are the parametric power transformations and a cubic spline family. Several data examples are presented to illustrate the methods.

Some Key Words: Errors-in-Variables; Nonlinear Models; Power Transformations; Regression Calibration; SIMEX; Spline Transformations; Transform-Both-Sides.

1 INTRODUCTION

Measurement error models concern the situation where one or more variables in a study cannot be measured exactly. We restrict our attention to the case where a single variable is measured with error. It is usually assumed that the relationship between the variable which is actually observed, W , and the true covariate of interest, X , is $W = X + U$, where U represents measurement error. Fuller (1987) applies this additive model for measurement error to many classical linear models.

There are also other ways to model the relationship between W and X , such as the multiplicative error model $W = Xe^U$, which gives additivity in the logarithmic scale, i.e., $\log(W) = \log(X) + U$. The idea behind both the additive and multiplicative error structure models is that, in the correct scale, measurement error is additive. The additive and the multiplicative error models are specific cases of a more general model $W = \mathcal{G}(X, U)$ for some function \mathcal{G} . In this article, we consider the set of functions \mathcal{G} such that $\mathcal{G}(X, U) = \mathcal{H}^{-1}\{\mathcal{H}(X) + U\}$, where \mathcal{H} is a monotone function with inverse \mathcal{H}^{-1} .

Additivity underlies almost all the measurement error models and modeling techniques in the common case that X is unobservable. The classical functional methods for ordinary regression (Fuller, 1997) and for general nonlinear models (Carroll, Ruppert & Stefanski, 1995) essentially without exception assume additivity. Likelihood (structural) methods which naturally allow for the commonly occurring within-person replication of the W 's typically assume additivity in some scale with a known distribution for U .

For all of these reasons, finding a scale for additive measurement error is important. In this paper, we investigate methods for determining an appropriate scale. Section 2 discusses two different methods for determining the correct scale for additivity of measurement error, the correlation method and the error distribution method. In section 3 we describe the transformations used, and in section 4 we describe their implementation. In section 5 we present data examples to illustrate the methods.

2 FUNCTIONAL TRANSFORMATIONS

In measurement error models, the literature makes a distinction between *classical functional* models, in which the values of unobserved true values of X_i , $i = 1, \dots, n$ are considered to be a sequence of unknown fixed constants, and *classical structural models*, in which the values of X are considered to be random variables. We believe that a more fruitful classification scheme is that of *functional modeling*, where no assumptions are made about the distribution of the X_i 's, and *structural mod-*

eling, in which parametric assumptions are made about the distribution of the unknown X 's. For a full description of functional versus structural modeling, see Carroll, et al. (1995, pp 144–145).

Additive error models assume that there is a monotone function $h(\cdot)$ such that

$$h(W) = h(X) + U, \quad (1)$$

where the random variable U is independent of X . There is an essential difference between our work and that typical in transformations, namely that in our case X cannot be observed so that without any additional information, $h(\cdot)$ cannot be identified. In practice, this extra information comes from replicating the W 's, so that (W_{ij}) is observed for $i = 1, \dots, n$ units and $j = 1, \dots, J$ replicates per unit. The resulting errors (U_{ij}) are assumed to be independent of X_i , although they may be correlated either given i or marginally.

The issue we address in this paper is that of estimating the transformation function $h(\cdot)$. We propose two different methods, both of which are truly functional modeling methods, in that they make no assumptions about the distribution of X , so that the methods are robust to the distribution of the predictor.

There are two general methods we propose, *correlation* methods and *error distribution* methods. These two methods are derived from the properties of the transformation model (1), as follows.

Property 1: Define the within-person mean $\overline{W}_i(h)$ and the within-person standard deviation $s_i(h)$ as

$$\overline{W}_i(h) = J^{-1} \sum_{j=1}^J h(W_{ij}); \quad s_i(h) = \left[(J-1)^{-1} \sum_{j=1}^J \{h(W_{ij}) - \overline{W}_i(h)\}^2 \right]^{1/2},$$

respectively. Under model (1), if the errors are symmetrically distributed, then $\overline{W}_i(h)$ and $s_i(h)$ are uncorrelated. Thus the correlation method selects the transformation $h(\cdot)$ so that the sample correlation for $\overline{W}_i(h)$ and $s_i(h)$ equal zero. Ruppert & Aldershof (1989), Box, Hunter, & Hunter (1978), and Solomon & Cox (1992) each mention correlation type methods in different contexts.

Property 2: For the correct transformation, the distribution of

$$h(W_1) - h(W_2) = U_1 - U_2 \quad (2)$$

does not depend on X . In particular, if the U 's are multivariate normal so too is (2). If (U_1, \dots, U_J) have a multivariate t -distribution with k degrees of freedom, then (2) is a multiple of a t -distribution with k degrees of freedom. If the U 's are independent with a mixture of two

normals distribution, (2) has a symmetric mixture of three normals distribution. These ideas suggest that the second class of methods, the error distribution methods, transform so that the terms in (2) follow one of the distributions mentioned. We address distributional shape via the Anderson-Darling (Anderson & Darling, 1954) and Filliben correlation (Filliben, 1975) statistics. There are times when a distributional model for the measurement error is desirable or even essential. Carroll, et al. (1995) describe several techniques which require that the measurement error distribution be specified, including Simulation-Extrapolation (Chapter 4), corrected scores (Chapter 6), conditional scores (Chapter 6), and likelihood techniques (Chapter 7).

3 TWO FAMILIES OF TRANSFORMATIONS

3.1 Definition of the Power Transformation

The power transformation family was described in Box and Cox (1964). The transformations in this family are indexed by the scalar parameter θ and have the form

$$h(v|\theta) = v^{(\theta)} = \begin{cases} (v^\theta - 1)/\theta, & \theta \neq 0 \\ \log(v), & \theta = 0. \end{cases}$$

Power transformations are monotone for each fixed θ . However, we have found that the restricted shape of power transformations limits their utility somewhat in our context, and for that reason we describe below an alternative family.

3.2 Definition Of The Spline Transform

The transformation family \mathcal{H} which we consider is the set of all zero-intercept, cubic piecewise-polynomial spline functions with knots at $\xi = (\xi_1, \dots, \xi_p)$. Transformations from this family have the form

$$h(v|\Theta, \xi) = \theta_1 v + \theta_2 v^2 + \theta_3 v^3 + \sum_{k=1}^p \theta_{k+3} (v - \xi_k)_+^3 \quad (3)$$

where $(a)_+^3 = a^3$ if $a > 0$, and $= 0$ otherwise. In general, the problem of picking knot points ξ is a difficult one. We will assume throughout this discussion that, given the data, the knot points ξ are fixed. For a more detailed discussion of knot point selection, see, for example, Eubank (1988).

Our method of selecting the knot points is as follows:

1. Let the combined data vector be \widetilde{W}^* , including all replicates across all units.
2. Let $\xi_1 = \widetilde{W}^*(0.005)$, where $\widetilde{W}^*(p)$ is the p th sample percentile of \widetilde{W}^*

3. Let $\xi_2 = \widetilde{W}^*(0.01), \xi_{p-1} = \widetilde{W}^*(0.99), \xi_p = \widetilde{W}^*(0.995)$
4. Let $\xi_i = \widetilde{W}^*(0.01 + \{p - 3\}^{-1}\{i - 1\}\{0.99 - 0.01\}), \quad i = 3, \dots, p - 2$

One should also note that the transformation given in (3) does not have the usual constant term, which is not identifiable. The parameter vector $\Theta = (\theta_1, \dots, \theta_{p+3})$.

The issue of obtaining a monotone transformation function is difficult. For a given set of data and knot points ξ , describing the set $\{\Theta : h(v + \delta|\Theta, \xi) - h(v|\Theta, \xi) \geq 0, \forall \delta > 0\}$ is a nontrivial analytical problem. Certainly, requiring that $\Theta_i \geq 0, i = 1, \dots, p + 3$ is sufficient to obtain a monotone transformation, but this is clearly unduly restrictive.

Our solution to the problem of obtaining a monotone transform is to create a set of $M + 1$ grid points $\{\tau_0 < \tau_1 < \dots < \tau_m\}$ on which we require that

$$h(\tau_i|\Theta, \xi) - h(\tau_{i-1}|\Theta, \xi) \geq 0, \quad i = 1, \dots, M. \quad (4)$$

We usually set $M = 23$; let τ_0 through τ_8 be the 1st, 2nd, 3rd, ..., and 9th percentiles of the combined data vector \widetilde{W}^* ; let τ_{15} through τ_{23} be the 91st, 92nd, 93rd, ..., and 99th percentiles of \widetilde{W}^* ; and let τ_9 through τ_{14} be set at evenly-spaced percentiles of \widetilde{W}^* between the 9th and the 91st, similar to step (4.) in the procedure describing the selection of knot points. We have found that through careful selection of these grid points, the final transformation will be monotone through the range of the data.

4 CORRELATIONS AND DISTRIBUTIONS

Using Property 1 of section 2, the correlation methods find the transformation which makes the within-person mean and standard deviation have zero sample correlation. With the power transformations restricted to the range $-3 \leq \theta \leq 3$, we have always observed a unique zero numerically, although in principle this need not be the case. The spline transformation has also had satisfactory numerical behavior, although there is no guarantee of a unique zero.

4.1 Assessing The Need For A Transformation Using Correlation and Powers

An important question to answer is whether or not the data indicate that a transformation would be appropriate. One way to answer this question is to create a confidence interval for θ_0 , the value at which the population correlation between the within-person mean and standard deviation equals zero. One rejects the “no transform” null hypothesis if the interval does not contain 1. We constructed an asymptotic confidence interval for θ_0 using the delta method and sandwich covariance

estimate via standard techniques, but while the confidence level for this interval is correct asymptotically, the convergence to the nominal level is quite slow. We also considered confidence intervals using resampling techniques described in Efron and Tibshirani (1993). Figure 1 shows the results of a simulation study comparing the asymptotic confidence interval to a confidence interval created using the bootstrap estimate of standard error. We studied other resampling-based confidence intervals with the same result: the delta-method confidence interval with sandwich covariance matrix estimate converges to its nominal level much more slowly than any of the bootstrap methods.

4.2 Assessing Distributional Shape

We consider the spline transformation to normality when there are exactly two replicates. The power transformations are even easier to work with. The overall goal is to find a vector $\hat{\Theta}$ which, for a given data set, makes the differences $E_i = h(W_{i1}|\hat{\Theta}, \xi) - h(W_{i2}|\hat{\Theta}, \xi)$ look as “normal” as possible, while satisfying the constraints given in (4). Actually, one need not specify a normal distribution for the measurement error. We investigate both bivariate t distributions and normal mixture distributions later in this article. There are several ways to check for normality of the differences E_i for a given value of Θ . We have chosen to use the probability plot correlation coefficient (PPC) described in Filliben (1975), which is a relative of the Shapiro-Wilk W statistic described in Shapiro and Wilk (1965). The basic idea is to calculate the correlation coefficient for a QQ-Plot of the E_i . The closer the empirical distribution of the E_i is to a normal distribution, the closer the PPC for the E_i should be to 1. Hence, our method of estimating Θ is to find the value $\hat{\Theta}$ which, subject to the constraints in (4), maximizes $\gamma(E)$, where $E = (E_1, \dots, E_n)^T$ and $\gamma(v)$ is the PPC for the vector $v = (v_1, \dots, v_n)^T$.

In this maximization problem, both the constraints and the objective function have simple matrix expressions. Given the data $\{W_{ij}\}$, $i = 1, \dots, n$, $j = 1, 2$ and a set of knot points ξ , define the matrices \mathbf{D} and \mathbf{C} as

$$\begin{aligned} \mathbf{D}_{ik} &= \begin{cases} W_{i1}^k - W_{i2}^k & i = 1, \dots, n \quad k = 1, \dots, 3 \\ (W_{i1} - \xi_{k-3})_+^3 - (W_{i2} - \xi_{k-3})_+^3 & i = 1, \dots, n \quad k = 4, \dots, p+3 \end{cases} \\ \mathbf{C}_{ik} &= \begin{cases} \tau_i^k - \tau_{i-1}^k & i = 1, \dots, M \quad k = 1, \dots, 3 \\ (\tau_i - \xi_{k-3})_+^3 - (\tau_{i-1} - \xi_{k-3})_+^3 & i = 1, \dots, M \quad k = 4, \dots, p+3 \end{cases} \end{aligned} \quad (5)$$

Thus, the maximization problem is to find $\max_{\Theta} \gamma(\mathbf{D}\Theta)$ subject to $\mathbf{C}\Theta \geq 0$, where by $\mathbf{C}\Theta \geq 0$ we mean that each element of $\mathbf{C}\Theta$ is nonnegative. The constrained maximization is accomplished using the FORTRAN program NPSOL (Gill, Murray, Saunders & Wright, 1986).

In modeling data such as the examples we discuss in Section 5, it is possible that the error

distribution may be something other than normal. We consider alternate distributions for the measurement error, specifically the bivariate t_k distributions (Johnson and Kotz, 1972) for $k = 20, 10, 8, 6, 4, 3$, and find separate transformations for each possible error distribution. Note that the bivariate t_k distributions are such that if $(U_1, U_2) \sim \text{Bivariate } t_k$, then $(U_1 - U_2)/\sqrt{2} \sim \text{Univariate } t_k$. The modification to the PPC statistic is simple—one calculates the correlation coefficient for the QQ-plot of the specified distribution instead of the normal distribution. As an additional check, for each transformation, we calculate the Anderson-Darling A statistic for the vector of differences E (Anderson & Darling, 1954).

We found with most of our examples that the spline transformation based on the error method transforms the data such that the error distribution is either normal or “nearly normal”, i.e., a bivariate t distribution with either 20 or 10 degrees of freedom, with the non-normality being attributable to a small number of points in the difference vector E . Another reasonable way to model the data is to assume that the measurement error is distributed as a two-component normal mixture distribution, with the measurement error for a (relatively small) number of data pairs being generated by a normal distribution with slightly heavier tails. We selected four normal mixture distributions, each chosen to have the same first four moments as a univariate t_k distribution, for $k = 20, 10, 8$, and 6 , respectively. We use the shorthand $\text{NM}(k)$ to refer to such a normal mixture distribution. For further information about the $\text{NM}(k)$ distributions, see the Appendix.

4.3 The Spline Transform With More Than 2 Measurements Per Individual

Unlike with correlations, the error distribution methods which model the distribution of the differences given in (2) do not have an easy direct definition for the case of $J > 2$. There are a variety of possibilities, including transformations so that the within-person sample standard deviation has the distribution of a sample standard deviation of a candidate error model, in which case the results of the previous subsection apply. Alternatively, one may wish to analyze the data pairwise, as this can often point out unusual replicates. Here we describe such a pairwise implementation.

In order to select the optimal Θ value, we must first determine the appropriate distribution for U , and then optimize with respect to that measurement error distribution. We select the distribution for U by some preliminary analyses on two columns of data. If the data are measurements on the same individual taken over time, then it makes some sense to use the two columns of data for which the measurements are farthest apart chronologically.

We implement the preliminary analyses in two stages. In the first stage, we select two columns of data, and find separate estimates $\hat{\Theta}_k$, $k = \infty, 3, 4, 6, 8, 10, 20$, where $\hat{\Theta}_\infty$ is the value of Θ which

maximizes $\gamma(E)$ for normally-distributed measurement error, and $\hat{\Theta}_k$, $k > 0$ is the value of Θ which maximizes $\gamma(E)$ for measurement error with a bivariate t_k distribution. For each value of $\hat{\Theta}_k$, we examine the PPC and AD statistics for the difference vector E , and for $\hat{\Theta}_\infty$ we also examine the PPC and AD statistics for the NM(k) distribution for $k = 6, 8, 10$, and 20. We also calculate the intra-individual mean/standard deviation correlation for the two selected data columns for each $\hat{\Theta}_k$.

Using the calculations in the first stage as a guide, we then select an appropriate $\hat{\Theta}_k$, say $\hat{\Theta}_*$ for additional analysis in the second stage. In this stage, we apply the transformation $h(v|\hat{\Theta}_*, \xi)$ to every data value W_{ij} , $i = 1, \dots, n$, $j = 1, \dots, J$, and then do an analysis of the differences of the transformation for each possible pair of columns. In this difference analysis, we calculate PPC and AD statistics, and their p-values, for the normal distribution, bivariate t_k distributions with $k = 3, 4, 6, 8, 10$, and 20, and NM(k) for $k = 6, 8, 10$, and 20. We also calculate the intra-individual mean/std correlation for each pair of columns of transformed data.

By combining the two stages of analysis, we can select an appropriate distribution for the measurement error U . We can then define quantities as follows to find an optimal $\hat{\Theta}$ value. Specifically, if $\gamma\{(v_1, \dots, v_n)^T\}$ is the PPC function for the specified error distribution, let

$$\begin{aligned} A_{ikm}(\Theta) &= h(W_{ik}|\Theta, \xi) - h(W_{im}|\Theta, \xi) \\ Q_{km}(\Theta) &= \gamma\left[\{A_{1km}(\Theta), \dots, A_{nkm}(\Theta)\}^T\right] \\ \bar{Q}(\Theta) &= \frac{2}{J(J-1)} \sum_{k=1}^{J-1} \sum_{m=k+1}^J Q_{km}(\Theta) \\ \tilde{Q}(\Theta) &= \text{median}\{Q_{km}(\Theta)\}, 1 \leq k \leq J-1, k < m \leq J \end{aligned}$$

We can then consider both $\bar{\Theta}$, the numerical maximizer of $\bar{Q}(\Theta)$, and $\tilde{\Theta}$, the numerical maximizer of $\tilde{Q}(\Theta)$.

5 EXAMPLES

5.1 Urinary Sodium Chloride Data

The Urinary Sodium Chloride data are discussed in Liu & Liang (1992). In a study attempting to relate the incidence of hypertension with urinary sodium, overnight urine samples were taken from 397 men on 7 consecutive nights. The data from days 1–6 were available to us. Because the data have a very high autocorrelation, we examined the data from days 1 and 6, which have the least correlation in the errors and hence presumably the most stable statistical properties.

Transform	Optimization Criterion	Mean/Std Correlation	Error Dist. Comparison	PPC p-value	AD p-value
Power ($\theta = 2.304$)	Correlation	0.00	Normal	0.967	0.801
Spline	Correlation	0.00	Normal	0.963	0.788
	PPC(Normal)	0.01	Normal	0.968	0.808

Table 1: Comparing transformations to different error distributions for the USC Data. The spline transformations used 8 knot points.

The estimated power transform from the correlation method was $\hat{\theta}_0 = 2.304$, with bootstrap confidence interval [1.520, 2.688], thus indicating the need for a transformation. We tested the differences of the power transformed data for normality, and found a PPC p-value = 0.967, and an AD statistic p-value = 0.801. In both cases, the null hypothesis is that the difference vector E has a normal distribution, with low p-values indicating non-normality. Hereafter we shall say that a data vector “passes” a given test (either PPC or AD) for a certain distribution if the P-value for the calculated statistic is greater than 0.10. Thus, the difference vector E from the power transformation “passes” both the PPC and the AD tests for normality.

Table 1 shows the results of the error distribution method using cubic splines for estimating the transform. Each row in the table gives the transformation, the criterion for optimization, the within-person mean and standard deviation sample correlation, the distribution under which the PPC and AD statistics are computed, and their corresponding p-values. One can see that the differences from either spline transformation clearly pass the PPC and AD tests for normality, with acceptably low within-person sample mean versus standard deviation correlation.

Figure 2 compares the correlation method power transformation and the error distribution method spline transformation. The circles in the graph represent percentiles of the data, from the 1st to the 99th. Each transformation has been standardized to the same scale. For this data set, the power transformation and the spline transformation were almost identical.

We repeated the analysis using all pairs of days and all six days together, and with one exception the answers were similar. The exception occurs for the pair of days (5,6), which seem to behave together quite differently from all the others. We have no explanation for this behavior.

5.2 Framingham Heart Study

The Framingham heart study measured various factors such as age, smoking habits, and blood pressure for 1,615 men aged 31-65, attempting to link these factors to the presence of coronary

Transform	Optimization Criterion	Mean/Std Correlation	Error Dist. Comparison	PPC p-value	AD p-value
Power ($\theta = 1.726$)	Correlation	0.00	Normal	< 0.005	< 0.005
			t_{10}	0.071	0.098
			NM(10)	0.098	0.384
Spline	Correlation	0.00	Normal	< 0.005	0.149
			NM(10)	0.165	0.211
			PPC(Normal)	-0.085	< 0.005
	PPC(t_{10})	-0.112	NM(10)	0.791	0.335
			t_{10}	0.979	0.461
			NM(10)	0.952	0.354

Table 2: Comparing transformations to different error distributions for the LSBP Data. The spline transformations used 12 knot points.

heart disease. The data we analyze here are two systolic blood pressure (SBP) measurements, the first of which is the average of two SBP measurements taken during a physical exam, and the second of which is the average of two SPB measurements taken at another physical exam two years later. We actually pretransform the data by analyzing $\log(\text{SBP} - 50)$, which is a modification of the transformation originally suggested by Cornfield (1962) and which we will designate as LSBP.

For the pretransformed LSBP variable, using the correlation method with power transformation we found $\hat{\theta}_0 = 1.726$ with 90% bootstrap confidence interval [1.113, 2.339], and 95% confidence interval [0.996, 2.455].

For the error distribution method using the spline transformation, Table 2 shows the usual statistics for the transformations of the LSBP data to additivity with various error distributions. There are a number of points to note. The power transformation using the correlation method results in differences which are non-normal and do not “pass” tests for the t -distribution with 10 degrees of freedom. The spline transformation using the correlation method does pass the t_{10} and NM(10) distribution tests. The spline transformation which attempts to fit a normal distribution to the differences is unsuccessful in doing so, at least with this number of knots. All of these calculations suggest that the errors are heavier-tailed than the normal distribution. The spline transformation under the error distribution method for the NM(10) distribution is shown in Figure 3.

Transform	Optimization Criterion	Mean/Std Correlation	Error Dist. Comparison	PPC p-value	AD p-value
None	N/A	-0.028	Normal	< 0.005	< 0.005
			NM(20)	0.031	< 0.005
			NM(10)	0.543	0.109
			t_{10}	0.765	0.401
Power ($\theta = 1.056$)	Correlation	0.00	Normal	< 0.005	< 0.005
Spline	Correlation	0.00	Normal	0.130	0.329
	PPC(Normal)	0.05	Normal	0.750	0.350

Table 3: Comparing transformations for the % Calories from Fat data.

5.3 CSFII Data

Our third example involves the Continuing Survey of Food Intakes for Individuals (CSFII) data set (Thompson, et. al, 1992). This data set contains information on nutrient intakes for 2,134 women. The data contain multiple measurements for each woman for a variety of daily dietary components such as vitamin A, vitamin C, amount of saturated fat, total calories, etc. Four measurements for each component were gathered for each woman. The first measurement was based on an extensive interview, and the subsequent three measurements were based on follow-up telephone interviews.

We analyze one dietary component from the CSFII data, percent calories from fat, by considering the second and fourth measurements for each woman in the study. We choose not to use the first measurement because it was gathered in a different manner than the last three. The power transformation using the correlation method yields an estimate $\hat{\theta}_0 = 1.056$, with bootstrap confidence interval [0.948, 1.164]. However, as is shown in Table 3, the differences of the no-transform model fail both tests for normality. The no-transform differences do pass both PPC and AD tests for t_{10} and NM(10) distributions.

The spline transformations with 5 knots both pass the normality tests, with acceptably low within-person sample mean and standard deviation correlation of the transformed data values is 0.055. The graph of the spline transformation using the error distribution method is given in Figure 4.

6 DISCUSSION AND CONCLUDING REMARKS

We have presented two methods for transforming the data to achieve additive measurement error. The correlation method transforms so that the sample correlation between the within-person mean

and standard deviation equals zero, while the error distribution method transforms so that differences have a specified distribution. Within each method we used power transformations and transformations based on cubic splines. A question which may arise is, “why not just transform the data to normality?” Such a method has been suggested by Nusser et. al. (1997), who also use power transformations and cubic splines. This method, which we call the *marginal* method, selects $h(\cdot)$ such that $h(W_{i1}), i = 1, \dots, n$ is approximately normally distributed. Thus, it transforms the *data* to normality instead of transforming the *errors* to normality. The marginal method with power transformation is in wide use in nutritional epidemiology.

There is no intrinsic reason that the marginal method must find the “right” or “wrong” answer. Indeed, in many examples marginal methods will yield transformations which pass both our correlation and error distribution criteria. One drawback of marginal methods which is important in measurement error modeling can be seen by once again considering the concepts of functional and structural modeling. The methods of transformation we have suggested are functional, by which we mean that they make no explicit assumptions about the distribution of the unobservable X . This makes sense in the context of measurement error models, because of the emphasis in that field of functional modeling to estimate regression parameters.

Unlike our methods, marginal approaches are explicitly structural, and can depend in a strong way on the distribution of X . For example, consider the case that no transformation is necessary, so that $W = X + U$, $h(v) = v$ and U is normally distributed. Marginal methods transform so that \underline{W} is normally distributed, and hence they will properly conclude that no transformation is necessary only if X is also normally distributed. This does not mean that marginal methods have no value, far from it, but only that one needs some care in employing them. As a noteworthy example of such care, in their applications Nusser, et al. also check what we call Properties 1 and 2 in section 2.

One point to keep in mind is that if there are $J = 2$ replicates, then plots of the within-person standard deviation versus the mean will have an odd shape if a significant number of W 's approach a lower bound. For example, if the lower bound is zero, and if $W_1 \approx 0$, then the standard deviation $\approx W_2/2^{1/2}$ while the mean is $\approx W_2/2$, so that the plot of the standard deviation against the mean will in effect be bounded by a line with intercept zero and slope $2^{1/2}$.

Finally, there is no guarantee that one can find a single transformation which will achieve additivity as measured by the correlation method with a normal or nearly normal error distribution as measured by the error distribution method. The Framingham data using power transformations

are a good example of this issue. Ruppert & Aldershof (1989) address this issue in their context, and suggest estimating parameters either as a weighted average of the correlation and error distribution methods, or by weighting their estimating equations. This is an interesting issue for further exploration.

ACKNOWLEDGEMENTS

This research was supported by the National Cancer Institute (CA-57030). Carroll's research was partially completed while visiting the Institut für Statistik und Ökonometrie, Sonderforschungsbereich 373, Humboldt Universität zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award. We are extremely grateful to Professor H. J. Newton for help in numerical optimization.

REFERENCES

- Anderson, T.W. & Darling, A. (1954). A test of goodness of fit. *Journal of the American Statistical Association* 49 765–769.
- Box, G. E. P. & Cox, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B* 26, 211–246.
- Box, G. E. P., Hunter, W. G. & Hunter, J. S. (1978). *Statistics for Experimenters*. Wiley, New York.
- Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Regression*. London: Chapman & Hall.
- Cornfield, J. (1982). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Federation Proceedings* 21, 58–61.
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.
- Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, Inc., New York.
- Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, 17, 111–117.
- Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons, Inc.
- Gill, P. E., Murray, W., Saunders, M. A., & Wright, M.H. (1986). *User's Guide for NPSOL (Version 4.0): A Fortran Package for Nonlinear Programming*. Stanford, California: Stanford University.
- Johnson, N. L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. New York: John Wiley & Sons.
- Liu, X. & Liang, K. Y. (1992). Efficacy of repeated measures in regression models with measurement error. *Biometrics* 48, 645–654.

- Nusser, S. M., Carriquiry, A. L., Dodd, K. W. & Fuller, W. A. (1997). A semiparametric transformation approach to estimating usual intake distributions. *Journal of the American Statistical Association*, to appear.
- Ruppert, D. & Aldershof, B. (1989). Transformations to symmetry and homoscedasticity. *Journal of the American Statistical Association* 84, 437–446.
- Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611.
- Solomon, P. J. & Cox, D. R. (1992). Nonlinear component of variance models. *Biometrika* 79, 1–11.
- Thompson, F. E., Sowers, M. F., Frongillo, E. A. & Parpia, B. J. (1992). “Sources of fiber and fat in diets of U.S. women aged 19–50: Implications for nutrition education and policy,” *American Journal of Public Health* 82, 695–718.

7 APPENDIX

7.1 Mixture Normals

If $X \sim t_k$ and $k > 4$, the first and third moments equal zero and the second and fourth moments are $EX^2 = k/(k-2)$ and $EX^4 = 3k^2/(k^2 - 6k + 8)$, respectively. A corresponding mixture normal density with the same first four moments is defined as follows. It has density

$$f(y) = \sum_{j=1}^2 (\pi_j / \sigma_j) \phi(y / \sigma_j),$$

where $\phi(\cdot)$ is the standard normal density function, $\sigma_1^2 = 1$, $\sigma_2^2 = (2k)/(k-4)$, $\pi_1 = k^2/(k^2 + 2k - 8)$ and $\pi_2 = 1 - \pi_1$.

7.2 Details Of The Algorithm

The following are the steps for optimizing the PPC statistic with respect to the coefficient vector Θ . Assume that we have data Y_{ij} , $i = 1, \dots, n$, $j = 1, 2$, a vector of knot points $\xi = (\xi_1, \dots, \xi_p)^T$, and a specified measurement error distribution U . We will use the notation Y_i , $i = 1, 2$ to denote the vector $(Y_{i1}, \dots, Y_{in})^T$, and $Y^* = (Y_1^T, Y_2^T)^T$. Define the matrices \mathbf{D} and \mathbf{C} as in (5).

1. Let σ_d be the theoretical standard deviation of $U_{i1} - U_{i2}$.
2. Let s_Y be the sample standard deviation of Y^* .
3. Define $W_{ij} = Y_{ij}/s_Y$.
4. Generate random values $\hat{\Theta}_m, \hat{\Theta}_m = (\hat{\Theta}_{m,1}, \dots, \hat{\Theta}_{m,p+3})^T$ for $m = 1, \dots, 1000$
 - (a) Let $\hat{\Theta}_{mi}, i = 1, \dots, p+3$, be independent Uniform[-1,1];
 - (b) Let s_{dm} be the sample standard deviation of the elements of $\mathbf{D}\hat{\Theta}_m$;
 - (c) Multiply each element of $\hat{\Theta}_m$ by the factor σ_d/s_{dm} ;

- (d) Test for $\mathbf{C}\hat{\Theta}_m \geq 0$. If any element of $\mathbf{C}\hat{\Theta}_m$ is negative, “throw out” $\hat{\Theta}_m$ and generate $\hat{\Theta}_{m+1}$ in step (a) above;
 - (e) Calculate $\gamma_m = \gamma(\mathbf{D}\hat{\Theta}_m)$, the PPC statistic.
5. Use the value of $\hat{\Theta}_m$ which gave the maximum γ_m as the starting value in the numerical optimization program NPSOL to find $\max_{\Theta} \gamma(\Theta)$

Coverage Probabilities for Theta = 0.5

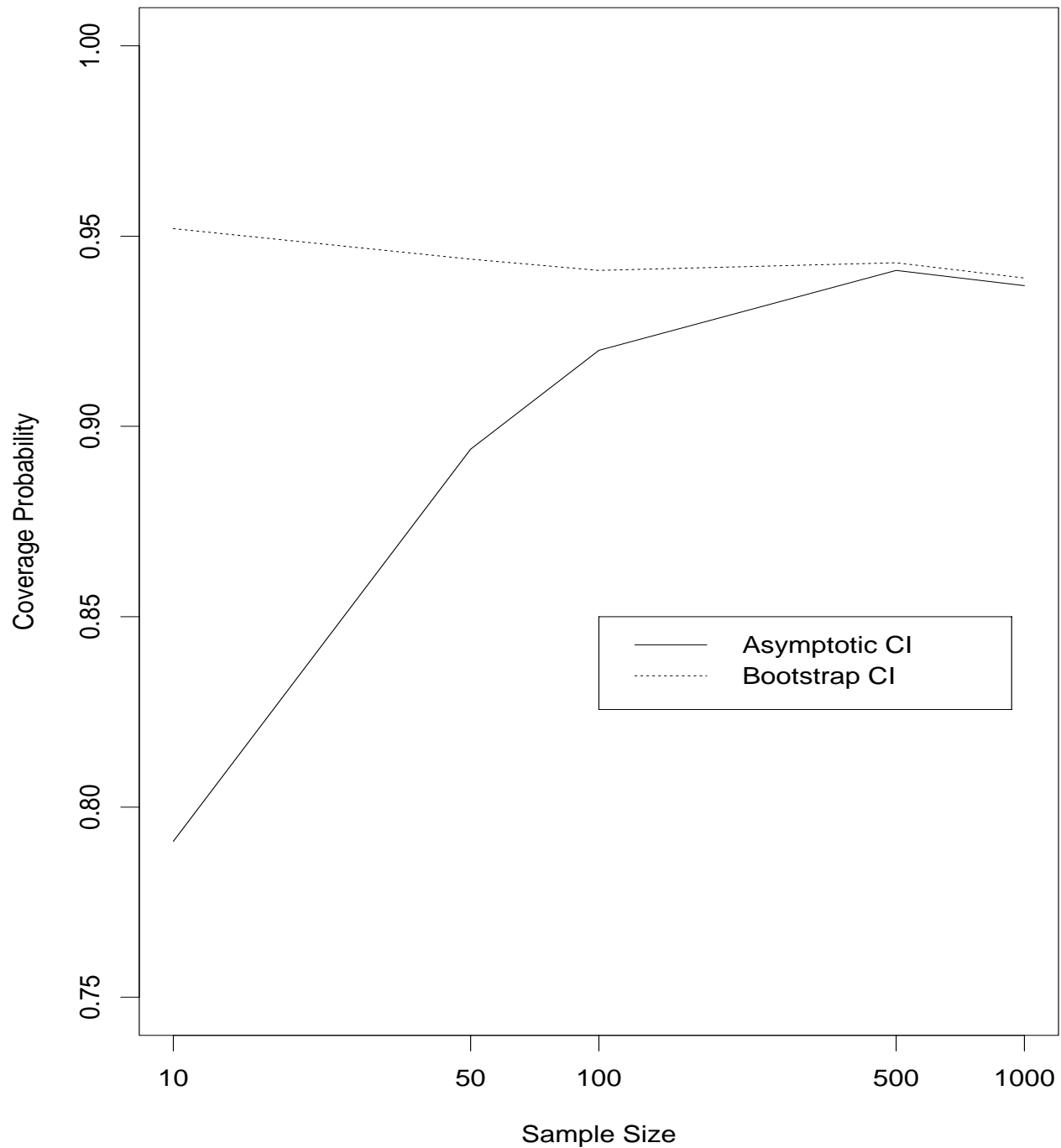


Figure 1: Coverage Probabilities for both the asymptotic confidence interval and the confidence interval which uses the Bootstrap Standard Error.

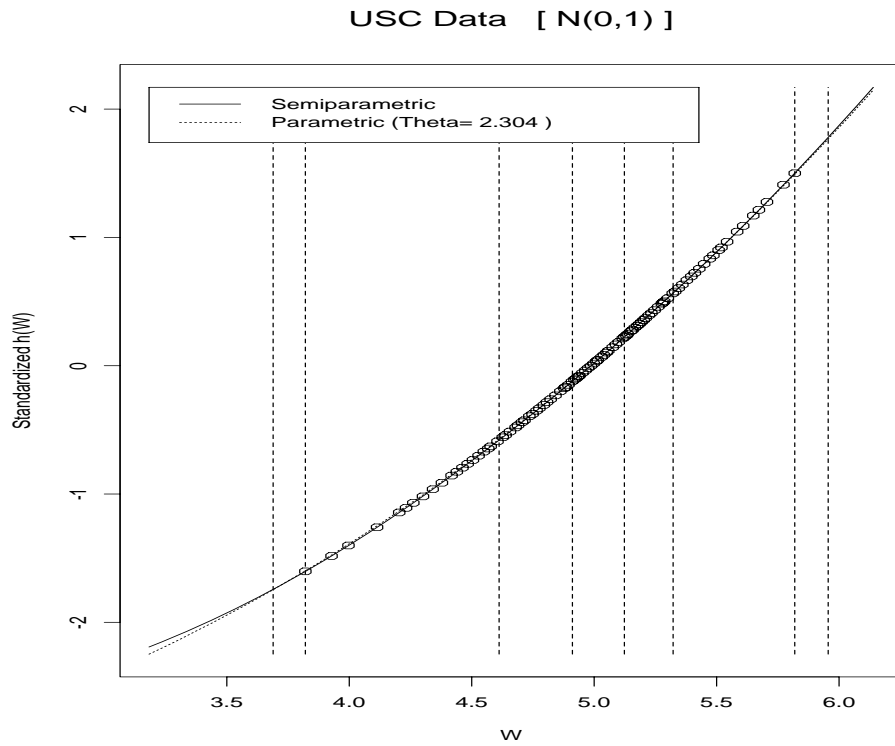


Figure 2: Graph of transformations for Urinary Sodium Chloride (USC) Data. The dashed vertical lines show the locations of the knot points.

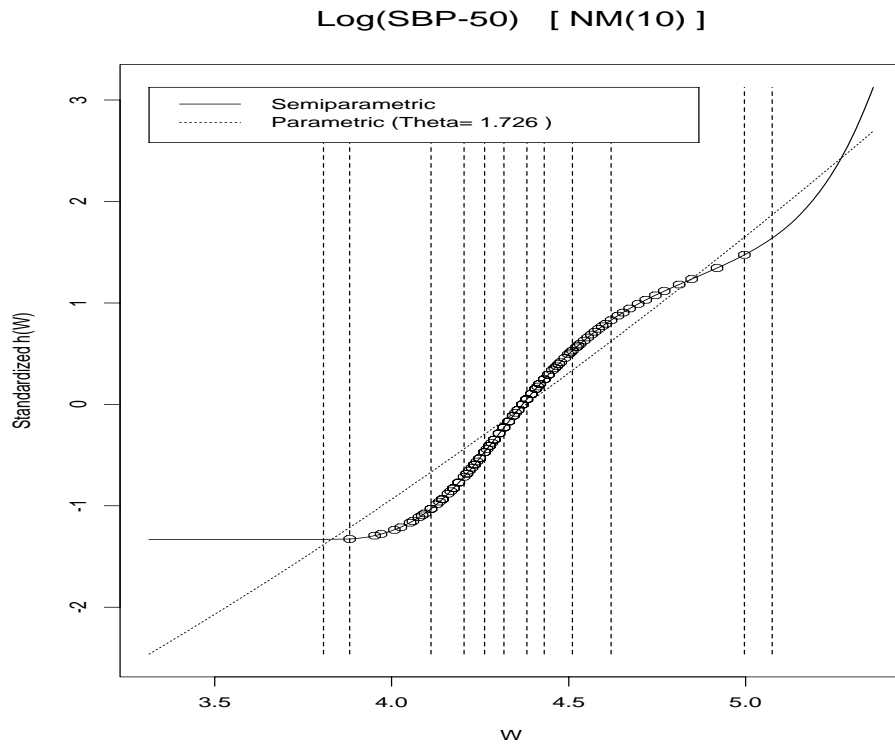


Figure 3: Graph of transformations for pretransformed systolic blood pressure (LSPB) data. The dashed vertical lines show the locations of the knot points.

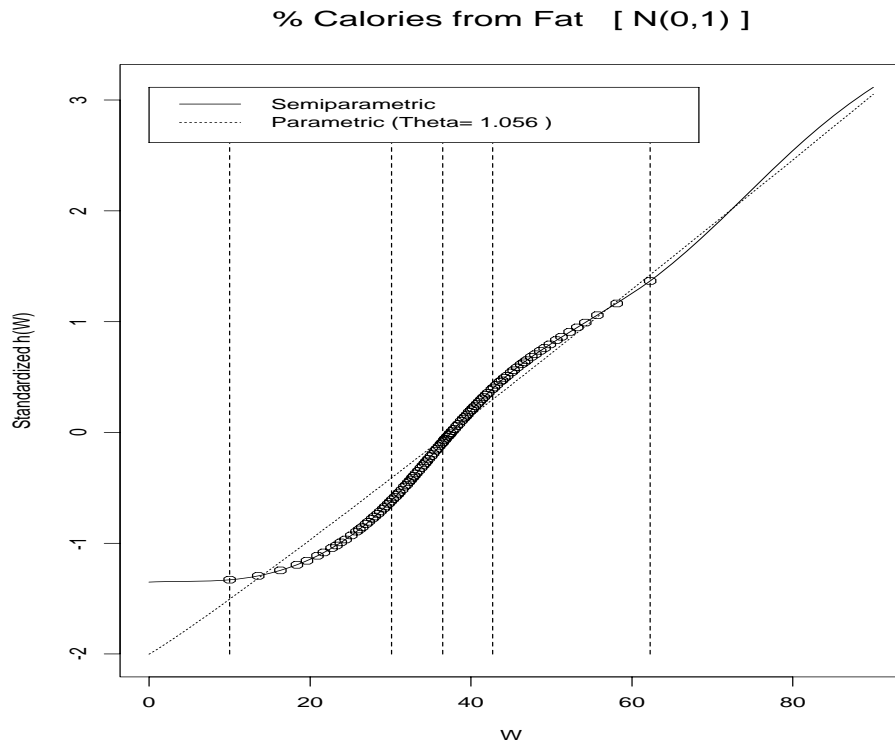


Figure 4: Graph of transformations for the CSFII % Calories from Fat (PCT) data. The dashed vertical lines show the locations of the knot points.